Clinical Tests to Diagnose Lumbar Segmental Instability: A Systematic Review

Low back pain (LBP) continues to be a major health problem and burden for individuals and society. As LBP is a heterogeneous condition, its classification into specific subgroups or syndromes has been suggested to aid the diagnosis of specific pathologies, assist in management decisions, and improve outcomes. Structural lumbar segmental instability (LSI) is universally recognized as an identifiable subgroup of individuals with LBP and is suggested to be a significant cause of morbidity associated with spinal dysfunction. The concept of structural LSI was first proposed by Knutsson, who advocated the assessment of LSI from the retrodisplacement (anterior-to-posterior translation) of lumbar vertebrae on lateral radiographs taken at end range spinal flexion and extension. Subsequently, White and Panjabi defined the related concept of “functional” lumbar instability as loss of the spine’s ability to maintain its pattern of displacement under normal physiological loads. Panjabi further described functional LSI in relation to the neutral and elastic zones of the functional spinal unit. Functional LSI is proposed to exist throughout spinal motion; but assessment is specifically focused on midrange spinal movements, where the neutral zone is suggested to be more manifest. While radiographic diagnosis of structural LSI is considered quantifiable, traditionally by assessment of vertebral translation at the end range spinal motion, the discrimination of functional LSI has not been consistently characterized in the literature and clinical diagnostic tests for functional LSI have not been specifically evaluated. It is, however, recognized that functional instability can exist in the absence of radiological evidence of LSI.

Cited causes of structural LSI include disc degeneration, postoperative spinal fusion (purported to produce abnormal

| STUDY DESIGN: Systematic literature review. |
| OBJECTIVES: To evaluate the diagnostic accuracy of clinical tests used to diagnose patients with structural lumbar segmental instability (LSI). |
| BACKGROUND: Patients with structural LSI represent an important, identifiable subgrouping of individuals with low back pain. Numerous clinical tests have been proposed to diagnose structural LSI; however, data on the diagnostic accuracy of these tests have not yet been evaluated through a systematic review of the literature. |
| METHODS: A systematic review was conducted in 6 electronic databases for diagnostic accuracy studies, published between January 1950 and March 2010, that evaluated clinical tests against radiological diagnosis of structural LSI. The diagnostic accuracy of the clinical tests from the retrieved articles was independently evaluated, reviewed, and quality scored using the QUADAS tool. |
| RESULTS: Four articles and a total of 11 clinical tests used in the diagnosis of structural LSI met the study inclusion criteria. The majority of tests had high specificity but low sensitivity, with positive likelihood ratios ranging from very small to moderate. QUADAS scores ranged from 16 to 25 out of a possible 26. The passive lumbar extension test was the most accurate clinical test, with high sensitivity (84%), specificity (90%), and a positive likelihood ratio of 8.8 (95% CI: 4.5, 17.3), indicating that this clinical test may be useful in the differential diagnosis of structural LSI. |
| CONCLUSION: This systematic review found that the majority of clinical tests routinely employed to diagnose structural LSI demonstrated only limited ability to do so. The results do, however, indicate that the passive lumbar extension test may be useful in orthopaedic clinical practice to diagnose structural LSI. Additional research is required to further validate its use for diagnosing structural LSI in all populations of those with low back pain. |


KEY WORDS: accuracy, low back pain, physical examination, validity

ABDULLAH M. ALQARNI, PT, MPhty • ANTHONY G. SCHNEIDERS, PT, PhD • PAUL A. HENDRICK, PT, MPhty
stresses on adjacent cephalad or caudal nonfused segments), postoperative disk excision or extensive decompression, and a history of trauma or recurrent LBP. Furthermore, a number of coexisting morphologies have been identified, including traction spurs, facet joint hypertrophy, and osteophytic formation. The estimated prevalence of structural LSI reported in the literature ranges from 12% of patients attending physical therapy for nonspecific LBP to 57% of patients with LBP who are suspected of having structural LSI.

The differential diagnosis of LBP subgroups, such as structural LSI, is considered vital for more effective management strategies. The diagnosis of structural LSI currently relies on radiographic confirmation, which exposes the patient to radiation and has limitations associated with access and cost. Effective noninvasive clinical tests to identify structural LSI would aid differential diagnosis, as well as improve understanding of the aetiology and management of this condition.

Numerous clinical examination findings have been proposed and promoted as signs and symptoms of structural LSI. These signs include palpation of vertebral malalignment, excessive passive intervertebral motion, subjective patient reports of “giving away,” “locking,” “through range pain,” an “instability catch” sensation during return from a flexed position, and a sensation of “slipping out” during the normal demands of spinal mobility. Additionally, subjective patient reports of pain exacerbation in the morning, on standing or rolling over, and with worsening weather, are all considered to be associated with structural LSI. However, to date, none of these clinical features have been validated as accurate diagnostic signs of structural LSI.

Orthopaedic clinical tests for structural LSI have also been routinely described in the literature and can be broadly divided into either passive or active tests. Passive tests include the posterior shear...
test, the prone instability test, passive accessory intervertebral motion (PAIVM) tests, passive physiological intervertebral motion (PPIVM) tests, and the passive lumbar extension (PLE) test. Active tests for structural LSI have included symptom reproduction during the sit-to-stand and the observation of an “instability catch” during return from flexion.

Despite the large number of clinical tests proposed to diagnose structural LSI, these tests have not yet been compared for their diagnostic accuracy and, consequently, no single test has been identified as superior to another and, subsequently, further investigated or incorporated exclusively into clinical practice. Therefore, the aim of this paper was to systematically review the literature related to clinical tests for structural LSI to establish which tests have the best diagnostic accuracy and utility in musculoskeletal and orthopaedic clinical practice.

**METHODS**

**Search Strategy**

A systematic search of relevant literature was conducted on November 1, 2009, and the search strategy results were monitored until March 1, 2010. A comprehensive search, with no language restriction, was conducted in the following databases: CINAHL, PubMed, MEDLINE, SCOPUS, AMED, and SPORTDiscus from January 1950 to March 2010. The following search terms were employed in various combinations, as outlined in the FIGURE: "low back pain," "instability," "lumbar spine," "spondylolisthesis," "anterolisthesis," "posteriorlisthesis," "spondyloysis," "validity," "sensitivity," "specificity," "diagnosis," "clinical tests," "provocation tests," "palpation," and "physical examination." There were no restrictions placed on the age of patients in the retrieved articles.

Studies considered for inclusion reported the use of clinical tests to diagnose structural LSI and were published as full reports before March 1, 2010. The inclusion criteria were that articles had to (1) report 1 or more clinical diagnostic tests for structural LSI, (2) establish a radiographic diagnosis of translational LSI (flexion-extension radiographs), and (3) report or allow computation of diagnostic accuracy (sensitivity, specificity, and positive and negative likelihood ratios) for the tests used to diagnose structural LSI.

From the results of the initial search, the first reviewer (A.M.A.) evaluated the titles and abstracts of retrieved articles for possible inclusion. Retained titles were subsequently assessed by 2 independent reviewers (A.M.A. and P.A.H.) for potential inclusion and retrieval of the full text article. Full text articles were screened independently for inclusion by 2 reviewers (A.M.A. and P.A.H.). If these 2 reviewers were unable to reach a consensus, a third reviewer (A.G.S.) was consulted. The reference lists of all included articles were searched for additional relevant references. The reviewers, who were experienced orthopaedic manipulative physical therapists and active researchers, were familiar with the literature and, therefore, not blinded to the authors, date of the publication, or journals in which the articles were published. Two international experts in LSI research were also consulted to ensure full inclusion of all potential articles on the diagnosis of LSI. A summary of the articles selected for review is presented in TABLE 1.

**Diagnostic Accuracy Statistics**

Articles investigating the diagnostic accuracy of clinical binary classification tests to diagnose radiographically confirmed structural LSI were required to report, or allow calculation of, sensitivity, specificity, and the positive likelihood ratio (+LR) and negative likelihood ratio (−LR) for each included test (TABLE 2).

**Quality Assessment**

The methodological quality of included articles was assessed independently by 2 reviewers (A.M.A. and A.G.S.), using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool developed by Whiting et al. The QUADAS tool is comprised of 14 items that are individually scored as either yes, no, or unclear (TABLE 3). Nine of the 14 items relate to bias, 3 to the quality of the reporting, and 2 to variability. The reviewers familiarized themselves with the QUADAS and discussed the quality items and the scoring system prior to the evaluation, to provide uniform interpretation of each study and to avoid quality assessment bias. When studies either satisfied or failed to meet the criteria, the 2 reviewers independently scored each of the 14 items yes or no, respectively. Items were scored as unclear when information lacked enough detail for the reviewers to decide whether the study satisfied or met a specific item. In the case of any disagreement, a third reviewer (P.A.H.) was consulted and adjudicated.

**Scoring and Quality of Papers**

The original QUADAS tool did not initially incorporate a system for scoring quality. For this systematic review, we used the methods subsequently proposed by the original developers. Item weightings, based and scaled for potential bias or variation, were used to develop the scoring system for the studies. Therefore, items 1, 5, 10, 11, and 12 were scored 3 points for yes, while items 3 and 6 were scored 2 points for yes, and all other items (2, 4, 7, 8, 9, 13, and 14) were scored 1 point for yes. All items were scored zero if the response was no or unable to be determined (unclear), which resulted in a total possible score of 26 (TABLE 3). As it was considered that the number of articles retrieved during this review would likely be low and it was recognized that rating scores might affect conclusions based on the quality of estimates of diagnostic accuracy, studies were not stratified into “high quality” or “low quality” using the QUADAS quality score.

**RESULTS**

The initial electronic database search yielded a total of 13,721 articles (FIGURE). After reviewing all
titles for key words and context, 230 articles were selected for possible inclusion in the review. After title duplications were removed, 80 article abstracts were screened, based on the inclusion criteria. After full text examination, 4 articles met the inclusion criteria and were vetted by 2 international experts who confirmed that they knew of no other published literature on this topic. All 4 articles were subsequently assessed using the QUADAS tool. A total of 11 clinical tests used in the diagnosis of structural LSI were reviewed from the 4 retrieved articles. The quality scores on the QUADAS tool ranged from 16 to 25 out of a possible 26. Sensitivity, specificity, likelihood ratios (LRs), and associated confidence intervals (95% CIs) were calculated from the study data if they were not specifically provided in the original articles. The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 1).

The study by Abbott et al had a QUADAS quality score of 19/26 and reported high specificity (89%) but low sensitivity (29%) for PAIVM (+LR, 2.5) to diagnose structural LSI in a cohort of 138 patients (mean age, 40 years) with CLBP, recruited from physical therapy clinics. PAIVMs in flexion were highly specific (99.5%) but showed low sensitivity (16%) for PAIVM (+LR, 2.5). Extension PAIVM also had low sensitivity (18%, 90%, and 1.9 for aberrant motions; 36%, 86%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for posterior shear test; 18%, 90%, and 1.9 for Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for prone instability test (TABLE 2).

The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 1).

The study by Abbott et al had a QUADAS quality score of 19/26 and reported high specificity (89%) but low sensitivity (29%) for PAIVM (+LR, 2.5) to diagnose structural LSI in a cohort of 138 patients (mean age, 40 years) with CLBP, recruited from physical therapy clinics. PAIVMs in flexion were highly specific (99.5%) but showed low sensitivity (16%) for PAIVM (+LR, 2.5). Extension PAIVM also had low sensitivity (18%, 90%, and 1.9 for aberrant motions; 36%, 86%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for posterior shear test; 18%, 90%, and 1.9 for Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for prone instability test (TABLE 2).

The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 1).

The study by Abbott et al had a QUADAS quality score of 19/26 and reported high specificity (89%) but low sensitivity (29%) for PAIVM (+LR, 2.5) to diagnose structural LSI in a cohort of 138 patients (mean age, 40 years) with CLBP, recruited from physical therapy clinics. PAIVMs in flexion were highly specific (99.5%) but showed low sensitivity (16%) for PAIVM (+LR, 2.5). Extension PAIVM also had low sensitivity (18%, 90%, and 1.9 for aberrant motions; 36%, 86%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for posterior shear test; 18%, 90%, and 1.9 for Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for prone instability test (TABLE 2).

The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 2).

The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 2).

The study by Fritz et al had a high (25/26) QUADAS quality score. The sensitivity, specificity, and +LR reported for tests to diagnose LSI in a cohort of 42 patients with chronic LBP (CLBP) were, respectively, as follows: 46%, 81%, and 2.4 for PAIVM; 57%, 48%, and 1.1 for the posterior shear test; 36%, 86%, and 2.5 for the Beighton hypermobility scale; 18%, 90%, and 1.9 for aberrant motions; and 61%, 57%, and 1.4 for the prone instability test (TABLE 2).
The study population consisted of 122 elderly patients (mean age, 69 years) with chronic back pain, who were not necessarily representative of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice. Nevertheless, the majority of studies did not provide a detailed description for the execution of the index test, sufficient information about the target condition, or the reference standard results interpreted without knowledge of the results of the index test.

The QUADAS tool identified study criteria that were considered to be important to the methodological quality of the retrieved articles. For instance, the study by Maing et al was judged to have a limited spectrum of patients, who were not necessarily representative of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al. The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.

The majority of studies did not provide sufficient information about the period between the performance of the clinical tests and the subsequent radiological examination, with the limitation of the spectrum of patients who would receive the test in practice, while the Kasai et al study lacked sufficient description for this criterion and, subsequently, received the same quality score. Patient selection criteria were clearly described by all articles, except Kasai et al; however, all studies used the same radiographic flexion-extension reference test to diagnose structural LSI, independent of the clinical index result. All clinical testing procedures were fully detailed to permit replication, except the aberrant motion test in the study by Fritz et al.
that the target condition might have changed in the interim.\textsuperscript{1,23,30,34} The blending of assessors for the clinical testing procedures and radiographic diagnosis of spinal instability was reported in all studies, as was the incidence of uninterpretable test results and withdrawal of patients. Abbott et al\textsuperscript{1} and Maigne et al\textsuperscript{33,34} did not report whether interpretation of the radiological examination occurred without knowledge of the clinical test results, while only Fritz et al\textsuperscript{23} reported on the knowledge and availability of full, relevant clinical data for interpretation of the clinical test results, as used in clinical practice.\textsuperscript{23}

The quality score for each item, total score, and the percentage agreement between the 2 reviewers is presented in TABLE 3. All items had an initial reviewer agreement that ranged from 50% to 100%. Items 1 and 10, which were related to patient spectrums and interpretation of the index test results retrospectively, had the lowest agreement (50%). Seven items (2, 3, 4, 6, 9, 11, and 12) had 75% agreement, while 5 items (5, 7, 8, 13, and 14) had a 100% agreement, between the 2 reviewers. The third reviewer (P.A.H.) adjudicated in all cases where agreement was not initially reached, resulting in the final scores reported in TABLE 3.

**Discussion**

The purpose of this systematic review was to evaluate the current evidence for clinical tests to diagnose structural LSI in musculoskeletal and orthopaedic clinical practice. The ability of clinical tests to diagnose structural LSI independent of radiographic investigations is considered important to expedite diagnosis and guide subsequent management, while limiting the exposure of patients to associated risks and further costs.\textsuperscript{58} A total of 11 clinical tests used in the diagnosis of 351 patients were identified from 4 articles that met the inclusion criteria for the study.\textsuperscript{1,23,30,34}

The results of the review show that diagnostic specificity values for all tests were consistently higher than sensitivity values, with the exception of the posterior shear test and prone instability test (TABLE 1). A negative result for a test with high sensitivity indicates that the test may have value in ruling out LSI, and a positive result for a test that has high specificity may be useful to rule in the condition.\textsuperscript{15} Likelihood ratios incorporate both the sensitivity and specificity of a test, and provide a direct estimate of how much a test result will change the odds of having the condition.\textsuperscript{15} In this study, the +LRs for structural LSI tests ranged from 1.1 to 8.8. The clinical interpretation of these values is presented in TABLE 2.

Two studies\textsuperscript{1,23} examined the ability of PAIVM\textsuperscript{23,35} to diagnose structural LSI. The study by Fritz et al\textsuperscript{23} reported low sensitivity (46%) and relatively high

---

**TABLE 4**

**Study Characteristics**

<table>
<thead>
<tr>
<th>Study</th>
<th>Subjects</th>
<th>Examiners</th>
<th>Definition of Instability From Flexion-Extension Radiographs*</th>
<th>Index Tests</th>
<th>Definition of Positive Index Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maigne et al\textsuperscript{33,34}</td>
<td>42 patients; mean ± SD age, 54.9 ± 9.8 y; CLBP greater than 2 mo</td>
<td>3 physicians</td>
<td>Translation greater than 10% (4.5-5.0 mm)</td>
<td>Sit-to-stand</td>
<td>Pain upon sitting down and relieved by standing up</td>
</tr>
<tr>
<td>Fritz et al\textsuperscript{23}</td>
<td>49 patients; mean ± SD age, 39.3 ± 11.3 y; LBP; median duration of symptoms, 78 d</td>
<td>1 physical therapist</td>
<td>Translation greater than 4.5 mm or greater than 15% of the vertebral body width</td>
<td>PAVM</td>
<td>Hypermobility</td>
</tr>
<tr>
<td>Abbott et al\textsuperscript{1}</td>
<td>138 patients; mean age, 40 y (range, 20-75 y); RCLBP; greater than 3 mo</td>
<td>27 physiotherapists</td>
<td>Translation greater than 2 SDs from the reference mean of asymptomatic individuals using Gaussian definition</td>
<td>PAVM</td>
<td>Symptom provocation</td>
</tr>
<tr>
<td>Kasai et al\textsuperscript{30}</td>
<td>122 patients; mean age, 68.9 y (range, 39-88 y); LBP; duration 1 mo to 5 y</td>
<td>3 orthopaedists; 2 examined PLE, 1 examined other tests</td>
<td>Translation motion of 5 mm PPIVM (extension)</td>
<td>PLE test</td>
<td>Hypermobility and/or pain</td>
</tr>
</tbody>
</table>

**Abbreviations:** LBP; low back pain; PAIVM, passive accessory intervertebral movements; PLE, passive lumbar extension; PPIVM, passive physiological intervertebral movements; RCLBP, recurrent chronic low back pain; ROM, range of motion.

*Reference tests for all studies.
†Aberrant motion include instability catch, painful arc, thigh climbing, or reversal of lumbopelvic rhythm.
For a diagnostic test to be valid, it must have acceptable reliability. The reliability of PAIVM testing is dependent on a number of intrinsic and extrinsic factors, and poor interrater reliability has been reported for judgment of segmental mobility. It is suggested that standardization of PAIVM testing would improve the reliability and, potentially, the ability of these tests to diagnose structural LSI. 

Abbott et al reported that flexion PPIVM had very high specificity (99.5%) but very low sensitivity (5%), and a moderate +LR (8.7) (95% CI: 0.6, 134.7) for the diagnosis of structural LSI. However, the width of the +LR confidence interval indicates the imprecision of the estimate, which was potentially due to an inadequate sample size to effectively evaluate the diagnostic accuracy of this test. The authors also found that extension PPIVM had low sensitivity (16%), high specificity (98%), and a moderate +LR (7.1) (95% CI: 1.7, 29.2) to diagnose LSI. These results suggest that extension PPIVM, compared to flexion PPIVM, might have greater clinical accuracy in the diagnosis of LSI, which is not unexpected, given the extension bias seen in the majority

### TABLE 5

<table>
<thead>
<tr>
<th>Test</th>
<th>Description of Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sit-to-stand&lt;sup&gt;24&lt;/sup&gt;</td>
<td>Pain immediately on sitting down and relieved by standing up.</td>
</tr>
<tr>
<td>PAIVM&lt;sup&gt;3&lt;/sup&gt;</td>
<td>Intervertebral motion was tested with the patient prone. The examiner contacted the spine with the hypothenar eminence and produced a posterior-to-anterior force. The mobility of each segment was judged as normal, hypermobile, or hypomobile. The presence of pain was recorded as present or absent.</td>
</tr>
<tr>
<td>Posterior shear test&lt;sup&gt;23&lt;/sup&gt;</td>
<td>The test was performed with the patient standing, with hands across the lower abdomen. The examiner placed one hand over the patient's cross arms. The heel of the other hand was placed on the patient's pelvis for stabilization. The examiner produced a posterior shear force through the patient's abdomen, and an anteriorly directed stabilizing force with the opposite hand. The test was repeated at each lumbar level by changing the position of contact of the posterior hand. A positive test occurred if familiar symptoms were provoked, and it is proposed to indicate lumbar instability.</td>
</tr>
<tr>
<td>Prone instability test&lt;sup&gt;23&lt;/sup&gt;</td>
<td>The test was performed with the patient prone, with the trunk supported on the examining table and the feet resting on the floor. With the patient in this position, the examiner performed a PAIVM test at each level of the lumbar spine. Any provocation of pain was recorded. The patient then lifted the feet off the floor and the PAIVM test was repeated. If pain was present in the resting position but subsided in the second position, the test was positive.</td>
</tr>
<tr>
<td>Aberrant motion&lt;sup&gt;27&lt;/sup&gt;</td>
<td>Any aberrant motions that are present during flexion-extension ROM, including instability catch, painful arc of motion, thigh climbing, reversal of lumbopelvic rhythm.</td>
</tr>
<tr>
<td>Beighton hypermobility scale&lt;sup&gt;21&lt;/sup&gt;</td>
<td>General ligamentous laxity was assessed with the 9-point Beighton scale. One point was given for each of the following: knee hyperextension greater than 10°, elbow hyperextension greater than 10°, fifth finger hyperextension greater than 90°, thumb abduction to contact the forearm, and ability to flex the trunk and place hands flat on the floor with knees extended.</td>
</tr>
<tr>
<td>PPIVM&lt;sup&gt;6&lt;/sup&gt;</td>
<td>PPIVMs were assessed with the patient sidelying, and consisted of moving the patient's spine through sagittal forward-bending (flexion) and backward-bending (extension), using the lower extremities, while palpating between the spinous process of the adjacent vertebrae to assess the motion taking place at each motion segment.</td>
</tr>
<tr>
<td>PLE test&lt;sup&gt;28&lt;/sup&gt;</td>
<td>The subject was in the prone position. Both lower extremities were then passively elevated, concurrently, to a height of about 30 cm from the bed, while maintaining the knees extended, and gently pulling the legs.</td>
</tr>
<tr>
<td>Instability catch sign&lt;sup&gt;29&lt;/sup&gt;</td>
<td>The subject was asked to bend his/her body forward as much as possible and then to return to the erect position. A subject who was not able to return to the erect position because of sudden low back pain was judged to have lumbar spinal instability.</td>
</tr>
<tr>
<td>Painful catch sign&lt;sup&gt;20&lt;/sup&gt;</td>
<td>The subject was asked to bend his/her body forward as much as possible and then to return to the erect position because of sudden low back pain was judged to have lumbar spinal instability.</td>
</tr>
<tr>
<td>Apprehension sign&lt;sup&gt;20&lt;/sup&gt;</td>
<td>The subject was asked whether he/she had felt a sensation of lumbar collapse because of sudden low back pain, when performing ordinary acts, including bending back and forth, from side to side, and sitting down or standing up. The subjects who had experienced such a sensation were judged to have lumbar spinal instability.</td>
</tr>
</tbody>
</table>

Abbreviations: LSI, lumbar segmental instability; PAIVM, passive accessory intervertebral movements; PPIVM, passive physiological intervertebral movements; PLE, passive lumbar extension; ROM, range of motion.
of both active and passive testing procedures for LSI.

The posterior shear test was originally described by Delitto et al. as a test to diagnose LSI. The results of this review demonstrated relatively poor sensitivity (57%), specificity (48%), and, consequently, a small +LR (1.1) for this test, which has also been shown to have poor intrarater (κ = 0.27) and interrater (κ = 0.22) reliability. These results indicate that the posterior shear test has limited overall diagnostic ability to diagnose LSI.

The prone instability test demonstrated low to moderate sensitivity (61%) and specificity (57%), and a low +LR (1.4), which suggests that the test has limited ability to accurately diagnose structural LSI. Moderate intrarater reliability (κ = 0.69) and the previously reported “almost perfect” interrater reliability (κ = 0.87) for this test are suggested to support this finding.

The Beighton hypermobility scale has been previously shown to have a positive correlation with structural LSI. This test had low sensitivity (36%) and high specificity (86%), and a relatively small +LR (2.5). The high intrarater reliability (ICC = 0.72) and specificity values in the retrieved study, combined with previously reported high reliability coefficients (ICC = 0.79), suggest that this test may be of some clinical use to rule in patients with a positive diagnosis of LSI. However, the low +LR, which in clinical interpretation is considered to be of greater diagnostic value than sensitivity and specificity values alone, makes this unlikely.

Maigne et al. examined the ability of the sit-to-stand test to diagnose structural LSI. The study compared patients whose LBP occurred immediately upon sitting down and was relieved on standing up, with a control group whose LBP did not show this pattern. The sit-to-stand test had a specificity of 100% and a sensitivity of 31% within this patient population. However, recruitment bias might explain the high specificity, as only patients who previously reported a positive sit-to-stand test were selected for the study group. This article had the lowest QUADAS score, and the authors reported that the test result might vary, depending on time of day that the test was conducted, the type of seat employed, and the patients’ symptom levels before the test. The authors also reported that these symptoms were observed in only 1 of 70 patients presenting with CLBP, which is much lower than the expected incidence of structural LSI in the CLBP population. Due to these factors, additional research is needed to address identified study limitations and to determine the test’s true diagnostic ability and clinical utility.

The PLE test is a relatively new method for examining structural LSI originally reported by Kasai et al. The test requires the patient to lie prone, while the clinician lifts both lower extremities into extension to a height of approximately 30 cm, while providing some traction to the lower extremities. During this maneuver, a positive test is based on an increase in pain that disappears on return to the starting position. In an attempt to diagnose structural LSI, Kasai et al. used the PLE test to examine 122 patients with mixed lumbar pathology. The PLE test was standardized and performed twice, at an interval of 2 to 4 weeks, by 2 independent orthopedists. They reported no disparate test results between the 2 sessions, which suggests substantial test-retest reliability. The sensitivity, specificity, and +LR of the PLE test were reported as 84%, 90%, and 8.8 (95% CI: 4.5, 17.3), respectively, indicating that the PLE test is a potentially effective clinical test to diagnose structural LSI. However, the prevalence rate for structural LSI in this study was relatively high (31%), which might be due to the study's elderly population sample, with high rates of spinal degeneration, stenosis, spondylolisthesis, and concurrent LSI. Therefore, although these results are promising, further investigation of this test should be undertaken in other patient populations, across different age groups, and with different assessors, to further evaluate its reliability and accuracy to diagnose structural LSI.

Kasai et al. also investigated a range of active movement signs/tests to diagnose structural LSI (TABLE 1). The instability catch sign, painful catch sign, and the apprehension sign all had relatively low sensitivity and high specificity, resulting in very small +LRs. These results suggest that these tests would more likely produce high false negative rates if used to diagnose structural LSI in research and clinical practice (TABLE 1). Similarly, Fritz et al. reported that a selection of aberrant motion test procedures demonstrated the same pattern of diagnostic accuracy and poor intrarater reliability (κ = −0.07), confirming the limited value of these signs/tests to diagnose structural LSI.

Clinical tests, such as those described in this review (TABLE 5), are not the only measures reported in the literature that are suggested to aid the diagnosis of structural LSI. Patient history of associated signs and symptoms suggestive of LSI has also been reported. Kasai et al. interviewed 368 patients with lumbar degenerative disease, of which 88 patients had structural LSI identified by imaging (translation greater than or equal to 5 mm). The results showed that pain on standing up and rolling over had the highest sensitivity (58% and 55%, respectively) and specificity (88% and 93%, respectively) for LSI, and a report of morning pain, with morning being the most painful time of day, had a sensitivity of 74% and specificity of 80%. Symptoms exacerbated by worsening weather had 65% sensitivity and 94% specificity. These results highlight the possibility that symptoms exacerbated by specific movements and the timing of symptoms could assist clinicians in diagnosing structural LSI.

The clinical implications of a positive diagnosis of structural LSI are inevitably related to either surgical or conservative management of this condition. Lumbar spinal fusion is usually suggested for...
patients with severe symptoms and radiographic evidence of hypermobility (greater than 4 mm of vertebral translation), who do not respond to conservative treatment. More commonly, conservative treatment is indicated; however, there are few studies that have specifically addressed the conservative management of radiologically determined structural instability. Reported conservative management has included braces or corsets and patient education to avoid overloading the passive stabilizing structures of the spine at end range. However, the mainstay of conservative management for instability-related lumbar spine pain has focused on exercises to improve neuromuscular control of the spine. Various strengthening programs targeting specific groups of muscles have been reported in the literature; however, no one specific program has demonstrated superiority over another. Further research is needed to identify the most effective strategies for patients with identified instability classifications. Conversely, the clinical implications of negative test findings for structural LSI are that alternative diagnoses need to be made and, in patients whose symptomology remains suggestive of lumbar instability, a potential diagnosis of functional LSI could be considered.

As with most diagnostic studies for lumbar pathology, limitations exist that may affect the validity of the results. Firstly, the literature that has reported the accuracy of clinical tests to diagnose structural LSI is limited, and only 4 of the articles retrieved met this literature review’s inclusion criteria. Secondly, all 4 articles used flexion-extension radiographs to identify abnormal vertebral translational motion to diagnose structural LSI. While flexion-extension radiographs have historically been considered the radiological reference test of choice for LSI, they have been suggested to be complicated by false positive rates and have significant variation in asymptomatic persons. Additionally, differences in patient positioning might account for the 10% to 15% variation in observed vertebral displacement.

The reported cutoff values for vertebral translatory motion employed to diagnose the presence of structural LSI also remain somewhat contentious and vary between 3 to 5 mm in the literature. The majority of studies in this review employed cutoff values for translational motion greater than 4.5 to 5.0 mm, with Abbott et al using translation beyond 2 SDs from the mean of an asymptomatic population as the cutoff value. The effects of differing translational cutoff values on the diagnostic ability of clinical tests to diagnose structural LSI are not currently known.

It has also been proposed that aberrant motion and dysfunction from structural LSI exist not only at end range but during midrange spinal movements, which these tests might not identify. Flexion-extension radiographs simply assess vertebral displacement statically at end range, which, theoretically, would only detect the function of the passive stabilizing subsystem. This might have significant limitations in detecting dysfunction from structural LSI that occurs within the neutral zone (midrange spinal motion).

Digital video fluoroscopy has also been utilized to identify normal and abnormal lumbar motion in vivo. This type of imaging has recently demonstrated an ability to identify movement abnormalities in patients with suspected functional LSI; however, further research is needed to confirm and substantiate these findings in a population with structural LSI. Despite these recognized limitations, flexion-extension radiographs remain, at present, the most common criterion reference standard to diagnose structural LSI.

A possible limitation of the methods used in this study was that only 1 reviewer searched the literature to identify articles for inclusion. Another factor that may limit the generalizability of the results of LSI diagnostic studies is the heterogeneity and sample size of the cohort under investigation. Study sample sizes identified in this review ranged from 42 to 138 patients (TABLE 4). These relatively low samples might have affected the internal validity and diagnostic accuracy of the results, and it has been suggested that studies of this nature should, theoretically, contain over 600 participants to have meaningful diagnostic value.

Additionally, quality assessment of retrieved articles is considered to be an essential component of most systematic reviews. The QUADAS tool was used to assess the quality of articles in this study; however, this and other well-utilized tools are suggested to have a number of associated limitations. These include the possibility that even well-conducted studies may score poorly if the methods and results of the study are not reported in sufficient detail, and that most quality assessment tools used in diagnostic studies do not include items that assess statistical power, which can subsequently affect a study’s validity.

Almost all of the clinical tests investigated in this systematic review, due to their high specificity, demonstrated the ability to diagnose patients with LSI, when the tests were positive. However, the trade-off for the majority of tests was the low sensitivity, which means that these tests may not be able to rule out people who test negative for structural LSI. Both PAIVM and PPIVM appear to have modest ability to diagnose structural LSI. The PLE test, however, had both the highest +LR (8.8) and lowest –LR (0.2) of the tests investigated, demonstrating a moderate but important role for both ruling in and ruling out structural LSI. These results suggest that the PLE test may be useful in clinical practice to diagnose structural LSI.

It is, however, important to note that the reliance on a single test in isolation is not usually recommended in musculoskeletal and orthopaedic clinical practice, and it is likely that a combination of valid and reliable tests, and the inclusion of patient-specific signs and symptoms, including historical elements, might fur-
ther assist clinicians in diagnosing and managing patients with LSI. Further investigation of these combined diagnostic factors is required in future studies that include larger patient populations, differing age ranges, and different assessors, to ensure the validity and diagnostic accuracy of the tests.

CONCLUSION

This is the first systematic review that has been conducted to identify the accuracy of clinical tests for diagnosing structural LSI. A total of 11 clinical tests were identified from the literature that met the study inclusion criteria. The reviewed articles were considered to be of sufficient quality to ascertain the diagnostic value of each test evaluated. The majority of tests had high specificity but low sensitivity. The PLE test was found to have the highest combined sensitivity and specificity, as well as the highest +LR, suggesting that in the absence of, or as an adjunct to, radiological imaging this test might be of use in musculoskeletal and orthopaedic clinical practice to diagnose structural LSI. However, additional research of the diagnostic accuracy of the PLE test across a range of patient populations with different assessors is recommended to further evaluate its validity to diagnose structural LSI.

ACKNOWLEDGEMENTS: We gratefully acknowledge the assistance of Masahiro Takenuma for translation of the Japanese articles.

REFERENCES


